Semantic Web technologies for digital preservation : the SPAR project

Emmanuelle Bermès
Bibliothèque nationale de France
Quai François Mauriac
75706 Paris Cedex 13
+33 153794240
emmanuelle.bermes@bnf.fr

Gautier Poupeau
Atos origin
18 Avenue d'Alsace
92926 Paris la Défense Cedex
+33 662038635
gautier.poupeau@atosorigin.com

ABSTRACT
The national library of France (BnF) is in the process of setting up a digital preservation repository, named SPAR (Système de Préservation et d'Archivage Réparti – Distributed Preservation and Archiving System). The infrastructure for the system, bought in 2005, is designed to support 1,5 petabytes of storage by 2014. The software components of the system are currently developped by Atos Origin. The design of the SPAR system is based on the major digital preservation standard, the OAIS model[1]. The architecture is composed of several modules connected via web services and based on open source components. One of the main components of the system is the data management module : it will use RDF data stored in a RDF triple store. We explain here why RDF is relevant for digital preservation and how it will be implemented in SPAR.

Categories and Subject Descriptors
H.4.m [**Information Systems**]: information systems applications – *Miscellaneous*

General Terms
Performance, Experimentation, Standardization.

Keywords
Digital preservation, metadata, OAIS, RDF, METS

1.THE CONTEXT
Digital preservation is not limited to storage and back-up : it involves complex strategies aiming at providing a trusted environment where digital objects can evolve along with the changes in technology, hardware and software environments. To manage these evolutions, strategies such as emulation and migration have to be proceeded. In order to do so, it is necessary to collect, store and manage all the information relevant to preserve a digital object through its lifecycle.

The OAIS proposes an information model which describes relevant meta information for preservation. Thus, in the OAIS model, a digital object is always associated with this information or metadata. The digital components (file, bitstream...) along with the associated metadata compose an **information package**.

In the functional perspective, the metadata is extracted from the information package, and stored in the "data management" entity which provides query and access functions.

The BnF has chosen METS[2], a well-known XML schema standard in the library community, as packaging format. A METS file is composed of various blocks of metadata, and a structural map which describes the components of the digital object. For each information package, all the metadata describing the object is embedded in a single XML file following the METS schema. In addition, some of the metadata is not embedded in the METS files for each information package, but loosely coupled : there is a link from the METS file to a reference information in XML managed by the system (e.g. file format descriptions).

2.THE ISSUE
The application of the OAIS information model raises the issue of how and when the metadata is going to be used. The metadata is intended to be used within the "data management" entity in the OAIS model, and the corresponding module in the SPAR system, for :

- **discovery** : to know what is in the archive
- **management** : to proceed to preservation actions such as monitoring, statistics, integrity checks, migrations, etc.

When designing a system for the long term, it is not possible to imagine all the queries that will be relevant in the future. Complex queries involve data formats, periods of time, events that have occured to a series of digital objects, software or human agents involved in the processes, etc. The flexibility of the data management is thus a key point in the development of SPAR and we had to take this into account when designing the indexation functions for the data management module.

All the relevant metadata being available in the METS files and in the reference information, we had to map them and index them. Four options were possible : a XML database, a relational database, a RDF triple store or a search engine.

A risk analysis taking into account implementation issues, functionnal opportunities and persistence in the long term, revealed the RDF triple store as the best candidate for managing metadata in this context :

[1] Open Archival Information System - ISO 14721 : http://public.ccsds.org/publications/archive/650x0b1.pdf

[2] Metadata Encoding and Transmission Standard : http://loc.gov/mets

- the mapping from XML to RDF was considered more relevant and evolutive than the mapping from XML to a relational database,
- the querying and access functionalities were richer than those provided by a search engine, thanks to expressiveness of a standardised query language SPARQL,
- the scalability and robustness was expected to be better than with a XML database, taking into account the amount of expected metadata to be handled by the system (2 billions triples after 2 or 3 years).

Regarding the latter, a benchmark was realized with Virtuoso[3] and 2 billions triples were generated with the the LUBM[4] ; the results of this prototype were satisfying and confirmed our choice.

3.THE SOLUTION

The mapping from the METS files to RDF required a modelling work on the information packages, in order to create an ontology. We considered that a litteral mapping from METS to RDF was unsufficiently persistent. We needed a data model independant from any implementation choice. Thus, we used the OAIS information model which provides this level of abstraction through a classification of the types of metadata that are needed for digital preservation:

- descriptive information
- representation information (includes information regarding the file formats)
- preservation description information (includes reference, context, provenance and fixity)
- structure information (information extracted from the METS structural map).

Within SPAR, the URIs that identify the classes and properties are created using the 'info' URI scheme[5] with a naming strategy based on the OAIS information model. Each URI is composed of:

- the 'info' URI scheme
- the naming authority (bnf/spar)
- a type of information (representation, reference, context, provenance, fixity, structure, agent)
- and the name of the class and property in the ontology.

Example : <info:bnf/spar/provenance#hasEvent> corresponds to the "hasEvent" property in the "provenance" ontology.

The structure of the URIs naturally led us to create one ontology for each category of metadata as defined by the OAIS. Thus the ontologies used within SPAR are not tied to the choice of METS and can be re-used in any context of digital preservation as long as the OAIS is used as a reference model. The ontologies include specific elements that have been created for the need of the project, but also existing elements from reference ontologies such as the Dublin Core metadata terms[6].

Our work on the SPAR project has shown the importance of metadata matters in a digital preservation project, and the relevance of RDF solutions to address the issues of flexibility and use of the metadata. We started the project with the idea that the main component of a digital preservation repository is the digital objects management system, in the case of SPAR the Fedora Commons framework[7]. Then we discovered that the data management part is also a major challenge. Although Fedora uses basic RDF data, we decided to build an independant data management module with more complex RDF handling functions. In the SPAR system, this module built using the Virtuoso triple store will be a key component for digital preservation strategies in the perspective of the long term evolution of the system.

4.REFERENCES

[1]Bermès, E., Dussert Carbone, I., Ledoux, T. and Lupovici, C. 2008. Digital preservation at National Library of France: a technical and organization overview. World library and information congress : 74th IFLA conference and council (Quebec, Canada, August 10-14 2008). http://www.ifla.org/IV/ifla74/papers/084-Bermes_Carbone_Ledoux_Lupovici-en.pdf

[2]Dubin, D., Futrelle, J., Plutchak, J., 2006. Metadata enrichment for digital preservation. Extreme markup languages (Montreal, Canada, 2006). http://www.idealliance.org/papers/extreme/proceedings/html/2006/Dubin01/EML2006Dubin01.html

[3]Pearce, J., Pearson, D., Williams, M., and Yeadon, S. 2008. The Australian METS profile – A journey about metadata. Dlib Magazine, 14, 3/4 (March/April 2008). DOI= doi:10.1045/march2008-pearce

[3] Openlink Software product

[4] http://swat.cse.lehigh.edu/projects/lubm/

[5] http://info-uri.info

[6] http://dublincore.org/documents/dcmi-terms/

[7] http://www.fedora-commons.org/